HW 1

Brunero Liseo
Sapienza Università di Roma, Italy

October 24, 2017

1. 1. Assume a Dirichlet process (DP) prior, $DP(M, G_0(\cdot))$, for distributions G on $\mathcal{X}$. Show that for any (measurable) disjoint subsets B_1 and B_2 of $\mathcal{X}$, $\mathrm{Corr}(G(B_1), G(B_2))$ is negative. Is the negative correlation for random probabilities induced by the DP prior a restriction? Discuss.

2. Simulation of Dirichlet process prior realizations. Consider a $DP(M, G_0)$ prior over the space of distributions (equivalently c.d.f.s) G on $\mathbb{R}$, with $G_0 = N(0, 1)$.

   Use both Ferguson's original definition and Sethuraman's constructive definition to generate (multiple) prior realizations from the $DP(M, N(0, 1))$ for fixed M with values ranging from small to large. In addition to prior c.d.f. realizations, obtain, for each value of M, the corresponding prior distribution of the mean functional

$$\mu(G) = \int t dG(t)$$

   and for the variance functional

$$\sigma^2(G) = \int t^2 dG(t) \{ \int t dG(t) \}^2.$$

   (Note that, because G_0 has finite first and second moments, both of the random variables $\mu(G)$ and $\sigma^2(G)$ take finite values almost surely; see Section 4 in Ferguson, 1973.)

   Finally, consider simulation under a mixture of DPs (MDP) prior, which extends the DP above by adding a gamma prior for M, that is $M \sim \mathrm{Gamma}(3, 3)$.

   Then, the MDP prior for G is defined such that, given M

$$G|M \sim DP(M, N(0, 1)).$$

   To simulate from the MDP, one can use either of the DP definitions given draws for M from its prior.

3. **Posterior inference for one-sample problems using DP priors.**

Consider data = $\{y_1, \ldots, y_n\}$, and the following DP-based nonparametric model:

$$\begin{aligned} Y_i|G &\sim \text{i.i.d.} G, \quad i = 1, \ldots n; \\ G &\sim DP(M, G_0) \\ G_0 &\sim N(m, s^2) \\ m, s^2 &\text{ and } M \text{ fixed.} \end{aligned}$$

The objective here is to use simulated data to study posterior inference results for G under different prior choices for M and G_0, different underlying distributions that generate the data, and different sample sizes. In particular, consider:

- two data generating distributions: 1) a $N(0, 1)$ distribution, and 2) the mixture of normal distributions,

$$0.5N(2.5, 0.5^2) + 0.3N(0.5, 0.7^2) + 0.2N(1.5, 2^2),$$

  which yields a bimodal c.d.f. with heavy right tail;
- sample sizes n = 20, n = 200, and n = 2000.

Use three values of M, say, 5 and 100.

Discuss prior specification for the DP prior parameters m, s^2. For each of the 6 data sets corresponding to the combinations above, obtain posterior point and interval estimates for the c.d.f. G and discuss how well the model fits the data. Perform a prior sensitivity analysis to study the effect of m, s^2, M on the posterior estimates for G.